Title: Obecność specjalistycznego słownictwa naukowego w dokumentach dostępnych w Internecie

Author: Arkadiusz Pulikowski

Citation style: Pulikowski Arkadiusz (2004). Obecność specjalistycznego słownictwa naukowego w dokumentach dostępnych w Internecie. W: D. Pietruch-Reizes, W. Babik (red.), "Usługi – Aplikacje – Treści w gospodarce opartej na wiedzy" (S. 238-245). Warszawa: Polskie Towarzystwo Informacji Naukowej

Arkadiusz PULIKOWSKI
Uniwersytet Śląski, KATOWICE

Obecność specjalistycznego słownictwa naukowego w dokumentach dostępnych w Internecie

Przeprowadzono badanie, którego celem było znalezienie odpowiedzi na pytanie, czy w Internecie są publikowane dokumenty naukowe operujące bardzo specjalistycznym słownictwem. Na podstawie baz danych Journal Citation Reports oraz Science Citation Index wygenerowano sześć dziedzinowych słowników frekwencyjnych. Terminy o najniższej częstości pochodzące z tych słowników kierowano następnie do wyszukiwarki Google. Badanie powtórzono czterokrotnie, co sześć miesięcy. Wyniki pozwalają na stwierdzenie, że Internet jest bardzo bogatym źródłem informacji naukowej. W jego zasobach można znaleźć dokumenty zawierające bardzo specjalistyczne słownictwo.

The presence of technical vocabulary in documents available on Internet

The paper discusses the results of a research aimed at finding whether there are documents published on the Net comprising highly technical vocabulary. On the basis of *Journal Citation Reports* and *Science Citation Index* six disciplines frequency dictionaries were generated. Next, the terms of the lowest frequencies coming from the dictionaries were put in Google search engine. The research was repeated four times every six months. The results let say that Internet is a rich source of scientific information. There is possible to find in its resources documents including very specialized technical vocabulary.

Osoby poszukujące w sieci informacji naukowych zadają sobie pytanie, czy specjalistyczna terminologia stosowana w uprawianej przez nich dyscyplinie naukowej jest obecna w Internecie. Przeprowadzone przez Autora badania miały na celu sprawdzenie, czy Internet daje możliwość odnajdywania dokumentów zawierających specjalistyczne słownictwo naukowe.

Praca badawcza miała następujący przebieg: Na podstawie samodzielnie opracowanych sześciu dziedzinowych słowników frekwencyjnych[1] wyselekcjonowano terminy o możliwie najniższej częstości występowania w publikacjach z wybranych dziedzin. Następnie, terminy te zostały skierowane do wybranego serwisu wyszukiwawczego – Google. Założono, że jeśli wyszukiwarka internetowa będzie w stanie odszukać tak dobrane słownictwo naukowe z poszczególnych dziedzin, będzie to oznaczać, że Internet może być cennym źródłem wiedzy specjalistycznej. Przyjęto również, że jeżeli będą odnajdywane dokumenty zawierające słowa rzadko występujące w słownictwie danej dziedziny, to na tej podstawie można wnioskować, iż terminy o większej częstości wystąpień też będą wyszukiwane.

[1] Słownik frekwencyjny to uporządkowany alfabetycznie zestaw słów (lub połączeń wyrazowych) wraz ze wskazaniem przy każdym słowie częstości jego użycia w tekście o określonej długości.

Wyszukiwane dokumenty były w głównej mierze stronami systemu WWW, a także innymi rodzajami plików zawierających tekst, jak txt, pdf, doc, rtf, xls, ppt, co wynikało z wyboru wyszukiwarki. Z uwagi na to, że około ¾ stron WWW dostępnych w Internecie jest napisanych w języku angielskim, zarówno wybrany serwis wyszukiwawczy, jak i terminologia użyta do jego sprawdzania były anglojęzyczne. Wybór języka angielskiego był tym bardziej uzasadniony, ponieważ jest on powszechnie uznawanym międzynarodowym językiem świata nauki.

Tworzenie dziedzinowych słowników frekwencyjnych

Podstawowym problemem w przeprowadzonych badaniach było znalezienie dziedzinowych słowników frekwencyjnych, na podstawie których miało być dobierane słownictwo o niskiej częstości występowania, kierowane później do wybranego serwisu wyszukiwawczego. Okazało się, że brak zarówno polskich, jak i anglojęzycznych słowników tego typu. Jedynym wyjściem było zbudowanie słowników frekwencyjnych we własnym zakresie. Aby tego dokonać, potrzebne były obszerne teksty naukowe w wersji elektronicznej. Elektroniczny format dokumentów był warunkiem koniecznym do wykonania ich automatycznej transformacji do postaci słowników frekwencyjnych. Jako źródło tekstów mogły posłużyć w zasadzie tylko abstraktowe lub pełnotekstowe bazy danych. Do stworzenia dziedzinowych słowników frekwencyjnych wykorzystano bazy danych Journal Citation Reports (JCR) oraz Science Citation Index (SCI), Obie bazy są dostępne na Uniwersytecie Śląskim w systemie InfoWare.

Baza danych Journal Citation Reports służy do oceny wartości czasopism naukowych, na podstawie analizy cytowań publikowanych w nich artykułów. Umożliwia wyszukiwanie czasopism o najwyższym wskaźniku cytowań zwanym Impact Factor, a co za tym idzie najczęściej wykorzystywanych przez naukowców. Analizowane cytowania pochodzą z ponad 6000 światowych czasopism naukowych, wydawanych przez 3000 wydawców w 60 krajach. JCR jest rocznikiem ukazującym się w dwóch wersjach: Science Edition, zawierającej dane z ponad 4500 czasopism z nauk ścisłych i technicznych oraz Social Sciences Edition, obejmującej dane z ponad 1400 czasopism z nauk społecznych.

Baza Science Citation Index gromadzi opisy bibliograficzne, streszczenia i cytowania zawarte w 3200 najważniejszych światowych czasopismach z zakresu nauk ścisłych. Istnieją również odrębne bazy cytowań dla nauk społecznych – Social Science Citation Index (SSCI) i kierunków humanistycznych – Arts and Humanities Citation Index. Do badań została wybrana baza SCI z dwóch powodów. Pierwszy – to większa liczba uwzględnionych czasopism (np. SSCI gromadzi artykuły z 1400 czasopism w porównaniu z 3200 z SCI). Drugi, ważniejszy, to zalety słownictwa nauk ścisłych – jest ono bardziej zwięzłe i jednoznaczne, niż terminologia nauk społecznych i humanistycznych. Jest też mniej interdyscyplinarne.

W bazie Journal Citation Reports każde czasopismo jest zakwalifikowywane do jednej lub kilku dziedzin wyróżnionych w bazie. Ta cecha JCR została wykorzystana w celu utworzenia listy periodyków z kilku wybranych dziedzin wiedzy. Były to: zoologia (zoology), geologia (geology), patologia (pathology), psychiatria (psychiatry), biologia (biology) i astronomia (astronomy & astrophysics). Z każdej dyscypliny były wybierane tylko te czasopisma, które przypisano do tej jednej dziedziny i żadnej innej. Wybór czasopism dla poszczególnych dyscyplin dokonano z wykorzystaniem bazy Journal Citation Reports, Science Edition z roku 2000.

Następny krok polegał na sprawdzeniu, które z wybranych czasopism dla poszczególnych dyscyplin były rejestrowane w bazie SCI[2]. Nazwy tych periodyków stały się podstawą do utworzenia instrukcji wyszukiwawczych operujących na polu bazy „Full journal title" (pełny tytuł czasopisma). Każda z instrukcji dotyczyła innej dziedziny i zawierała inny zestaw tytułów czasopism. Przykładowa treść instrukcji wyszukiwawczej dla geologii (najkrótsza) przedstawia się następująco:

GEOLOGIA: (CARBONATES-AND-EVAPORITES OR ECLOGAE-GEOLOGICAE-HELVETIAE OR GEOLOGY OR JOURNAL-OF-GEOLOGY OR JOURNAL-OF-METAMORPHIC-GEOLOGY OR JOURNAL-OF-SEDIMENTARY-RESEARCH-SECTION-A-SEDIMENTARY-PETROLOGY-AND-PROCESSES OR JOURNAL-OF-SEDIMENTARY-RESEARCH-SECTION-B-STRATIGRAPHY-AND-GLOBAL-STUDIES OR SCOTTISH-JOURNAL-OF-GEOLOGY OR SEDIMENTARY-GEOLOGY OR SEDIMENTOLOGY)

Instrukcje były kierowane do dwóch baz, będących kumulacjami roczników SCI za lata 1991–1995 i 1996–2000, co w sumie daje okres dziesięcioletni. Efektem działania każdej z instrukcji wyszukiwawczych była lista opisów artykułów z czasopism dla konkretnej dziedziny wiedzy. Następnie otrzymane rekordy zgrywano na dysk twardy lokalnego komputera, uwzględniając tylko zawartość abstraktów. Dzięki takiemu postępowaniu otrzymano obszerne pliki tekstowe zawierające abstrakty z czasopism z poszczególnych dziedzin z lat 1991–2000. Tabela nr 1 prezentuje rozmiar otrzymanych plików oraz liczbę czasopism, jaka była brana pod uwagę na poszczególnych etapach ich selekcji w bazach JCR i SCI.

Tabela 1

Dane liczbowe opisujące proces tworzenia plików tekstowych zawierających abstrakty artykułów z czasopism

	Wszystkie czasop. z danej dziedziny rejestrowane w JCR	Czasop. zaliczone tylko do tej jednej dziedziny w JCR	Czasopisma odnalezione w SCI	Rozmiar pliku z abstraktami [MB]
Astronomia	37	28	24	76,9
Biologia	51	23	15	14,8
Geologia	36	18	9	9,3
Patologia	67	17	13	26,1
Psychiatria	82	19	13	16,2
Zoologia	112	61	22 (47)[3]	16,6

Otrzymane pliki stały się podstawą do stworzenia dziedzinowych słowników frekwencyjnych. Wygenerowano je przy użyciu programu napisanego w języku Perl. Jest to narzędzie programistyczne przeznaczone specjalnie do obróbki tekstu. Zawiera funkcje zoptymalizowane pod kątem operacji wykonywanych na dokumentach. Właściwym środowiskiem pracy języka programowania Perl są systemy operacyjne typu Unix, Linux. Na szczęście istnieją również

[2] Było to konieczne, ponieważ SCI nie indeksuje wszystkich czasopism zawartych w JCR. Pełną listą dysponuje tylko SCI Expanded, która w czasie prowadzenia badań nie była jeszcze dostępna dla Autora.

[3] W przypadku zoologii, z uwagi na nieproporcjonalną do innych dyscyplin liczbę czasopism znalezionych w bazie SCI (47), została ona zredukowana o przeszło połowę, z ograniczeniem do tych, w których tytule pojawiała się nazwa dziedziny – zoologia (22).

wersje napisane dla innych systemów operacyjnych. Program przetwarza plik z abstraktami i na jego podstawie tworzy plik wyjściowy zawierający nie posortowany słownik frekwencyjny. Alfabetyczne sortowanie według słów wykonano w programie Microsoft Excell. Podczas tworzenia słowników przyjęto zasady:

- elementami słownika są pojedyncze wyrazy;
- wyraz brany pod uwagę to ciąg małych lub wielkich liter oddzielonych znakami odstępu lub spacjami nie będącymi literami;
- jeżeli wybrany do słownika wyraz zawiera wielkie litery, to są one zamieniane na małe.

Tabela nr 2 pokazuje, ile było zliczonych wszystkich słów podczas tworzenia danego słownika (co jest równoznaczne z określeniem wielkości próby) oraz ile spośród tych słów to różne słowa tworzące słownik frekwencyjny. Jak można zauważyć, liczba słów utworzonego słownika nie zależy tylko od liczby wszystkich słów pliku wejściowego. Widać to najwyraźniej po porównaniu danych liczbowych słowników frekwencyjnych z astronomii i zoologii.

Tabela 2

Liczebność słownictwa poszczególnych słowników frekwencyjnych

	Wszystkie słowa	Wielkość słownika frek.
Astronomia	11 752 690	46 669
Biologia	2 141 496	55 798
Geologia	1 340 261	30 824
Patologia	3 631 969	45 249
Psychiatria	2 190 844	27 917
Zoologia	2 391 744	74 353

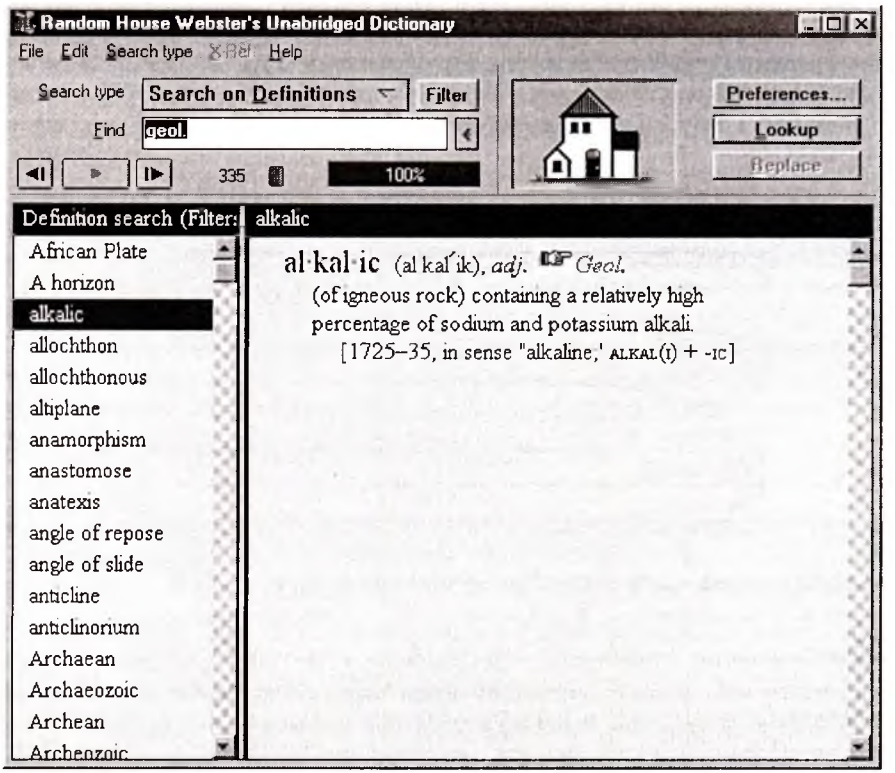
Metoda wyboru terminów ze słowników frekwencyjnych

Ponieważ Autor nie będąc ekspertem z danej dziedziny nie mógł ocenić, czy wyraz ze słownika frekwencyjnego utworzonego dla niej należy do niej, czy też do dyscyplin pokrewnych, trzeba było znaleźć inny sposób weryfikacji przynależności słów słownika do określonej dziedziny. W tym celu wykorzystano słownik języka angielskiego Random House Webster's Unabridged Dictionary, dostępny w wersji elektronicznej na dysku CD-ROM w Czytelni Oddziału Informacji Naukowej BUŚ. Indeks tego słownika zawiera 210 801 pozycji, co świadczy o ogromnym zasobie słownictwa jakie obejmuje. Omawiany słownik pochodzi z 1996 roku. Słownictwo specjalistyczne, zawarte w słowniku Webstera, jest odpowiednio oznaczane. Określenie dziedziny, do jakiej należy dany termin, pojawia się albo przy opisie jednego z jego znaczeń, albo przy haśle głównym. W celu wyeliminowania wieloznaczności pod uwagę brano tylko terminy opisane drugą z wymienionych metod. Rysunek nr 1 przedstawia przykładowe hasło z dziedziny geologii.

Niestety, system pomocy słownika Webstera nie dostarcza wykazu skrótów stosowanych do oznaczania w słowniku dyscyplin naukowych. Zostały one odszukane metodą prób i błędów. Znalezione dziedziny, wraz z odpowiadającą im liczbą terminów słownika, dla których oznaczenie dziedziny występuje w haśle głównym, zebrano w tabeli nr 3. Tabela uwzględnia tylko oznaczenia dyscyplin nauk ścisłych.

W tym momencie można już wytłumaczyć, dlaczego słowniki frekwencyjne tworzone były dla takich, a nie innych dziedzin. Wybór dyscyplin był kompromisem pomiędzy nazwami

dyscyplin wyróżnianymi przez słownik Webstera, a stosowanymi do kwalifikowania czasopism w bazie JCR. Niektóre bardzo rozwinięte dyscypliny, indeksowane w JCR, były dzielone na poddyscypliny, co uniemożliwiało stworzenie dla nich słowników frekwencyjnych. Barierą była tu też ogromna liczba czasopism. Przykładami takich dziedzin mogą być fizyka i chemia. Z kolei matematyka i informatyka w słowniku Webstera posiadają terminologię w większości wielowyrazową. Nie można jej niestety pogodzić z pojedynczymi słowami występującymi w tworzonych automatycznie słownikach frekwencyjnych. Mimo ograniczenia wyboru można jednak powiedzieć, że sześć wybranych dyscyplin dobrze reprezentuje nauki ścisłe.



Rys. 1. Przykład hasła słownikowego z dziedziny geologii

Proces weryfikacji słownictwa słowników frekwencyjnych został połączony z wyborem terminów o niskiej częstości wystąpień, przydatnych z punktu widzenia prowadzonego badania. Sprawdzano częstość kolejnych terminów słownika Webstera dla wybranej dziedziny w jej słowniku frekwencyjnym. Wybierane były słowa o najniższej częstości wystąpień – równej jeden na całą próbę. Trzeba tu zaznaczyć, że częstość wyrażona przez liczbę jeden nie oddaje w żadnym razie w sposób dokładny realnej częstości występowania słowa tak oznaczonego. Można jedynie uznać, że w badanej próbie słowo to miało najniższą możliwą częstość. Jak się okazało, wiele terminów słownika Webstera nie pojawiło się w ogóle w dziedzinowych słownikach frekwencyjnych. Może to świadczyć o ich jeszcze niższej częstości, ale niekoniecznie. Gdyby wielokrotnie zwiększyć wielkość próby, wówczas mogłoby się okazać, że nieodnalezione terminy miałyby podobną częstość do tych, które w omawianych słownikach mają częstość równą jeden. Ta sama uwaga odnosi się do słów

trochę częstszych od tych oznaczonych w słowniku przez jeden. Ponieważ z punktu widzenia prowadzonego badania nie jest istotna dokładna wartość częstości terminu, a jedynie to czy jego częstość jest niska, przyjęto kryterium dopuszczające do dalszych prac te terminy słownika Webstera, które albo nie wystąpiły w słowniku frekwencyjnym, albo uzyskały wartość częstości równą jeden dla całej próby. Przeważnie są to słowa o podobnie niskiej częstości. Z każdej dziedziny wybrano 50 takich terminów. Wyjątek stanowi astronomia, dla której z uwagi na liczne terminy wielowyrazowe, udało się zebrać tylko 22 słowa.

Tabela 3

Oznaczenia nazw dyscyplin stosowane w słowniku Webstera

Oznaczenia nazw dziedziny	Liczba terminów
chem.	2874
pathol.	1731
physics	845
pharm.	793
bot.	761
math.	747
biochem.	717
biol.	641
astron.	490
med.	472
zool.	467
geol.	335
computers	211
psychiatry	158

Wybór serwisu wyszukiwawczego

Do wyszukiwania w Internecie informacji bardzo szczegółowych wykorzystuje się wyszukiwarki lub metawyszukiwarki. Z uwagi na ograniczenie liczby wyświetlanych wyników, metawyszukiwarki zostały odrzucone jako narzędzie badawcze.

Oprócz wyszukiwarek ogólnego przeznaczenia brano pod uwagę zastosowanie wyszukiwarek specjalizujących się w wyszukiwaniu informacji naukowych. Aby dokonać wyboru przeprowadzono test pięciu wyszukiwarek naukowych i trzech ogólnych o największej liczbie indeksowanych stron. Wyniki zebrano w tabeli 4.

Z każdej z sześciu dziedzin wybrano po dwa (pierwsze z listy) terminy i skierowano je do kolejnych wyszukiwarek. Wśród wyszukiwarek naukowych tylko Scirus potrafił sprostać postawionemu zadaniu. Jego wyniki są jednak znacznie gorsze od osiąganych przez wyszukiwarki ogólne. Niekwestionowanym liderem okazał się Google i to on został wybrany do prowadzenia dalszych badań.

Wyniki badań

Badanie główne przeprowadzono 16 listopada 2001 r. Do wyszukiwarki Google skierowano kolejno wszystkie przygotowane dla poszczególnych dziedzin terminy. Łącznie

było ich 272 (5*50+22). W tabeli nr 5 zebrano wyniki w formie uproszczonej. W kolumnach tabeli znajduje się podzielona na przedziały liczba trafień dla danej dziedziny. Wartości w wierszach wyrażają procentowo i liczbowo sumę trafień dla danego przedziału.

Tabela 4

Wyniki testu pięciu wyszukiwarek naukowych i trzech największych wyszukiwarek ogólnych

	First Search	Cora	STN Easy	Sci Central	Scirus	Google (ang.)	Fast (ang.)	WiseNut (ang.)
altiplane	0	0	0	0	4	39	29	19
anamorphism	0	1	0	0	61	369	157	142
abiogenesis	5	0	0	0	324	5790	2624	3485
abiogenic	0	0	0	0	117	641	388	394
abetalipoproteinemia	0	0	0	0	154	1350	874	1063
abiotrophy	0	0	0	0	21	330	198	166
acalculia	0	0	0	0	108	890	488	557
acarophobia	0	0	0	0	12	502	336	341
achernar	6	0	0	0	224	4360	1902	1943
ananke	0	0	0	0	576	4660	3162	6557
abranchiate	0	0	0	0	5	137	54	124
acaudal	0	0	0	0	8	169	31	390

Tabela 5

Suma trafień w różnych przedziałach dla poszczególnych dziedzin

	0-100	**101-500**	**501-1500**	**1501 i więcej**
Astronomia	9% (2)	32% (7)	14% (3)	45% (10)
Biologia	40% (20)	34% (17)	8% (4)	18% (9)
Geologia	30% (15)	34% (17)	22% (11)	14% (7)
Patologia	24% (12)	38% (19)	24% (12)	14% (7)
Psychiatria	24% (12)	32% (16)	36% (18)	8% (4)
Zoologia	54% (27)	38% (19)	6% (3)	2% (1)
Razem	32% (88)	35% (95)	19% (51)	14% (38)

Badanie powtórzono czterokrotnie w odstępach półrocznych. Ostatnie miało miejsce w listopadzie 2003 roku. Dwa lata wcześniej – w listopadzie 2001 roku, w czasie pierwszego badania, liczba stron w bazie danych wyszukiwarki Google wynosiła ponad 1,6 mld. W czasie prowadzenia ostatniego badania w listopadzie 2003 roku było to już 3,3 mld. Można by oczekiwać, że skoro liczba indeksowanych stron wzrosła ponad dwukrotnie, to będzie to miało bezpośrednie przełożenie na liczbę wyszukiwanych dokumentów. Przypuszczenie to potwierdziło się. Jak się okazało średni wzrost liczby znalezionych dokumentów dla poszczególnych terminów mieścił się dla wszystkich dziedzin w przedziale od 2 do 3 razy. Tabela nr 6 przedstawia zmiany w liczbie trafień dla części badanych terminów z geologii.

Tabela 6

Zmiany w liczbie trafień dla części badanych terminów z geologii

Lp.	**Termin**	**XI '01**	**V '02**	**XI '02**	**V '03**	**XI '03**
1.	gumbotil	40	56	56	57	68
2.	gyttja	601	2330	2900	2990	4250
3.	hogback	8600	1200	13000	14000	18900

ciąg dalszy tabeli 6

4.	hornito	628	1390	1670	1810	2690
5.	hypocenter	12200	15100	20400	25600	23600
6.	interstade	150	205	247	312	427
7.	isocline	1090	1960	2210	2590	2590
8.	isoseismic	86	139	189	141	1970
9.	katamorphism	30	56	61	52	150
10.	laccolith	2570	3710	4160	3360	6780
11.	metacryst	46	84	76	14	56
12.	microseism	401	613	861	890	3110
13.	neocene	206	353	879	541	1090
14.	nivation	294	479	482	574	805
15.	oolith	120	411	627	527	668

Wnioski

Z przeprowadzonego badania wynika, że korzystając z internetowych serwisów wyszukiwawczych można znaleźć dokumenty zawierające bardzo specjalistyczną terminologię. Tylko w kilku przypadkach liczba wyszukanych stron była mniejsza niż dziesięć. W pozostałych – trafień były dziesiątki, setki a nawet tysiące. Liczba wyszukanych dokumentów dla większości terminów oscylowała w przedziale od kilkudziesięciu do kilkuset. Jest to z pewnością dobry rezultat, świadczący o ogromnym potencjale wiedzy gromadzonym w Internecie. W przypadku terminów, dla których odnaleziono niewielką liczbę dokumentów (do kilkudziesięciu) warto zastosować jedną z metawyszukiwarek. Podnosi to w sposób znaczący kompletność wyszukiwania. Trafność przeważającej większości wyników nie budzi zastrzeżeń. Takie są wyniki obserwacji. Trzeba mieć jednak na uwadze fakt, że słownictwo będące przedmiotem formułowanych zapytań jest bardzo specyficzne. Tak specjalistyczne terminy nie mogły się pojawić na stronach WWW i w innych typach dokumentów indeksowanych przez wyszukiwarkę Google zupełnie przypadkowo. Większość dokumentów stanowią prace naukowców i studentów. Można wśród nich znaleźć artykuły lub ich streszczenia, materiały konferencyjne, programy konferencji, bibliografie dorobku naukowego, materiały dydaktyczne przygotowane dla studentów, opisy projektów badawczych, prace zaliczeniowe i magisterskie, a nawet dysertacje doktorskie.

Poszczególne dziedziny nie mogą być ze sobą porównywane z uwagi na różną wielkość tekstów, jakie były podstawą utworzenia słowników frekwencyjnych z wybranych do badania dziedzin. Przekłada się to na różnice w częstościach słów (tych wybranych do badania) poszczególnych dyscyplin. Mimo to otrzymane wyniki są w dużym stopniu do siebie zbliżone.